



江苏常熟地区汉族人群47个微单倍型的遗传多态性及遗传结构分析*

潘坤鹏^{1,2)**} 冯耀森^{2)**} 于文帅^{2,3)} 刘宗伟⁴⁾ 姚伊人²⁾ 赵杰²⁾ 康克莱²⁾

张驰²⁾ 王乐^{2)***} 吴坚^{1)***}

(¹⁾ 昆明医科大学法学院, 昆明 650500; (²⁾ 公安部鉴定中心, 法医遗传学公安部重点实验室, 现场物证溯源技术国家工程实验室, 北京 100038; (³⁾ 中国人民公安大学侦查学院, 北京 100038; (⁴⁾ 江苏省常熟市公安局, 常熟 215500)

摘要 **目的** 探索江苏常熟地区汉族人群在47个常染色体微单倍型基因座的遗传多态性, 评估应用效能及法庭科学参数。**方法** 采用MHSeqTyper47混合DNA鉴定试剂盒进行基因座复合扩增及文库构建, 使用MiSeq FGx测序平台进行测序, 测得的数据应用MHTyper数据分析软件进行分析, 对获得的样本遗传信息进行评估, 结合千人基因组数据(1000 Genomes Project phase 3, 1KG)评估群体间遗传分化指数及遗传距离, 并计算法庭科学参数。**结果** 江苏常熟地区汉族人群与1KG中的中国北京群体的遗传分化和遗传距离最小, 并得到最接近的有效等位基因数(A_e), 累积随机匹配概率(combined matching probability, CMP)与1KG中东亚参考人群的5个群体均接近, 为 1.25×10^{-36} , 累积非父排除概率达0.999 999 999 964 1。**结论** 本研究报告了47个微单倍型基因座在江苏常熟地区汉族人群中的等位基因频率及遗传多态性信息, 为47个微单倍型在法医学应用中提供了数据基础。另外, 比较了1KG参考人群与江苏常熟地区汉族人群的多态性差异, 并揭示了47个微单倍型在江苏常熟地区汉族人群中的遗传结构。总的来说, 1KG中的东亚人群参考数据更符合江苏常熟地区汉族人群的遗传特征。

关键词 法医遗传学, 微单倍型, 二代测序

中图分类号 R89, D919

DOI: 10.16476/j.pibb.2023.0027

微单倍型是近年来国际法医遗传学界广泛关注的一类新型遗传标记, 被定义为200 bp或300 bp的DNA片段内, 至少两个单核苷酸多态性(single nucleotide polymorphism, SNP)位点组成的序列多态性遗传标记^[1]。由于微单倍型多SNP的特点, 往往表现出比单个SNP遗传标记更高的多态性^[2]。而且, 不同于传统短串联重复序列(short tandem repeat, STR)遗传标记, 微单倍型在DNA复制过程中不会因重复序列的复制滑脱现象从而干扰DNA分型^[3], 因此, 微单倍型被认为有助于混合DNA分析^[4]。

随着二代测序(next generation sequencing, NGS)技术的发展, 微单倍型遗传标记在法医遗传学领域中的研究逐渐广泛^[5-8]。微单倍型在多种场景的应用潜能被研究者们发掘, 如祖先推断^[9]、混合DNA解析^[10]、亲缘关系推断^[11]等。为了更好地评估微单倍型的应用效能, 研究者们公布了不

同微单倍型组合在各自研究人群中的频率和法医学参数等信息^[12-13]。但目前基于中国汉族人群微单倍型的群体结构分析仍较少, 评估微单倍型遗传多态性或遗传结构的研究多是基于公开访问的千人基因组(1000 Genomes Project, 1KG)数据。该数据所包含中国地区样本数据仅301个, 由中国南方群体(Southern Han Chinese (CHS), $n=105$)、中国北京群体(Han Chinese in Beijing, China (CHB), $n=103$)和中国西双版纳的傣族群体(Chinese Dai in Xishuangbanna (CDX), $n=93$)三

* 国家自然科学基金(81971797), 国家重点研发计划(2022YFC3341001)和公安部技术研究计划(2017JSYJA05)资助项目。

** 并列第一作者。

*** 通讯联系人。

王乐 Tel: 010-83752707, E-mail: wangle_02@163.com

吴坚 Tel: 0871-65922837, E-mail: wujianky@126.com

收稿日期: 2023-01-21, 接受日期: 2023-04-18

个群体^[14-15]组成。同时,低覆盖度的数据很可能导致等位基因的丢失^[16]。因此,有必要对中国汉族人群微单倍型深度测序进行遗传多态性和遗传结构分析。

本研究将选用针对案发现场常见的低质量DNA而设计的MHSeqTyper47微单倍型检测试剂盒^[16]。该试剂盒已整合的47个微单倍型基因座由Kidd等^[17]以及Chen等^[18-19]前期报道,其中37个微单倍型在中国汉族人群中被证实混合拆分效能较高^[13]。另外10个基因座具有良好的低质量DNA检测能力^[16],并具有理想的有效等位基因数(effective number of alleles, A_e)值。使用试剂盒对223名中国江苏常熟汉族无关个体的47个微单倍型基因分型。统计了取样群体的等位基因频率,描述了江苏常熟汉族人群的遗传多态性特征。本研究还对比了47个微单倍型基于1KG数据的理论效能和实际取样人群数据的差异,并基于Nei氏遗传距离和分化指数探索常熟汉族人群和1KG参考人群的遗传结构。最后基于取样人群的法庭科学参数评估了试剂盒的效能。

1 材料与方法

1.1 DNA提取及定量

223份江苏常熟地区汉族无关个体唾液卡样本由志愿者签署知情同意后提供,经过公安部物证鉴定中心科研伦理委员会审查后批准,批准号为2019-002。剪取大约25 mm²唾液卡,使用M48磁珠提取纯化试剂盒(Qiagen,德国)提取DNA。使用Qubit[®] 3.0荧光计对DNA进行定量,将定量后的模板DNA稀释至1 mg/L用于扩增。

1.2 文库制备及测序

文库制备按照MHSeqTyper47试剂盒(公安部鉴定中心,北京)用户手册进行两步PCR,第一步利用多重PCR扩增子捕获技术,扩增目标区域。第二步以连接有唯一标签和测序接头的标签引物扩增第一步PCR产物,获得测序平台兼容的文库。取1 ng DNA与10 μ l反应混合物1、4 μ l MH引物、补无核酸酶水(Thermo Fisher Scientific,美国)混合得到20 μ l PCR1反应混合物。反应混合物在95 $^{\circ}$ C预变性5 min,然后95 $^{\circ}$ C 30 s、60 $^{\circ}$ C 2 min、72 $^{\circ}$ C 2 min进行22个循环,72 $^{\circ}$ C 5 min,最后4 $^{\circ}$ C保存。取25 μ l反应混合物2、1 μ l标签7、1 μ l标签5、1 μ l PCR1产物与22 μ l无核酸酶水混合得到50 μ l PCR2反应混合物。反应混合物在95 $^{\circ}$ C预变性

5 min,然后95 $^{\circ}$ C 30 s、60 $^{\circ}$ C 2 min、72 $^{\circ}$ C 2 min进行15个循环,72 $^{\circ}$ C 5 min,最后4 $^{\circ}$ C保存。使用MHSeqTyper47试剂盒中包含的纯化磁珠纯化PCR2产物,纯化后文库在7500 PCR仪器(Thermo Fisher Scientific)上使用KAPA文库定量试剂盒(Roche,瑞士)进行定量。将纯化后文库均一化至6 nmol/L并混合到一起,取5 μ l混合文库使用5 μ l的0.2 mol/L NaOH(Sigma-Aldrich,美国)变性5 min,并使用HT1(Illumina,美国)终止变性和文库稀释。最后,使用MiSeq[®] Reagent v2试剂盒(Illumina)对600 μ l含有90%的12.5 pmol/L文库和10%的12.5 pmol/L PhiX Control V3(Illumina)的混合物进行测序。

1.3 数据分析

通过微单倍型分析软件MHTyper^[20]对测序结果进行分析,使用最低测序深度为10,基因座内等位基因测序深度和基因座内最高等位基因测序深度比值0.1作为阈值过滤可能来源于污染或噪声的测序读取序列。

分析得到的微单倍型序列通过MHTyper软件分配的数字等位基因进行命名,这将兼容下游法医学参数计算及群体遗传学分析软件。命名流程参考Song等^[21]的研究,即首先将组成一个微单倍型的SNP按照其在染色体上的位置进行排序,然后根据dbSNP数据库(<https://www.ncbi.nlm.nih.gov/SNP/index.html>)列举出每个SNP的等位基因,接着将SNP的所有等位基因进行组合得到所有可能的微单倍型等位基因,最后将微单倍型的所有等位基因组合按照字母表排序,并以阿拉伯数字序号作为等位基因的数字命名。

使用数字等位基因通过Alrequin v3.5软件进行Hardy-Weinberg平衡以及连锁不平衡检验^[22],使用PowerStats v1.2电子表格软件计算法庭科学参数^[23]。使用python 3编写的脚本计算 A_e 和累积随机匹配概率(combine matching probability, CMP)。线性回归分析由线上软件SPSSPRO(<https://www.spsspro.com/>)完成。成对遗传分化指数(fixation index, F_{st})与Nei氏遗传距离使用数字等位基因通过GenAlEx v6.5计算并通过MEGA Version 11.0.13绘制分子进化树。

2 结 果

2.1 测序参数和样本数据质量

223份样本使用3张PE MiSeq[®] V2流动池

(Illumina) 完成测序, 每张芯片分别检测 73、73 和 77 个样本。3 次测序中的平均簇通过率分布在 77.90% 至 90.21% 之间 (表 1), 与同类测序相

近^[24-25], 而 3 次测序的 Q30 比例分布在 72.49% 到 84.76% 之间。3 张芯片均得到了足以用于后续分析的读取序列数。

Table 1 Sequencing metrics of three runs for 223 unrelated individuals from Changshu, Jiangsu						
Sequencing No.	Number of libraries	Sequencing chips	Cluster pass rate/%	%≥Q30	Phas/Prephas	
					Read 1	Read 2
1	73	V2	85.76±3.56	84.76	0.094/0.017	0.285/0.124
2	73	V2	77.90±3.62	72.49	0.125/0.019	0.390/0.130
3	77	V2	90.21±0.75	75.75	0.054/0.004	0.577/0.117

排除人工检查判断为测序失败和污染的样本, 最终 216 份样本的数据用于后续数据分析。单个样本总序列数分布在 11 004 到 402 195 之间, 其中, 91.67% ($n=198$) 的样本总序列数在 40 000 到 200 000 之间。单样本平均有效序列比例分布在 0.931 到

0.992。图 1a 展示不同总序列数区间的样本数目及总序列数与有效序列比例的关系, 可以观察到有效序列比例随着总序列数而升高。各基因座 216 个样本的平均等位基因覆盖比 (allele coverage ratio, ACR) 分布在 0.835 到 0.912 (图 1b)。

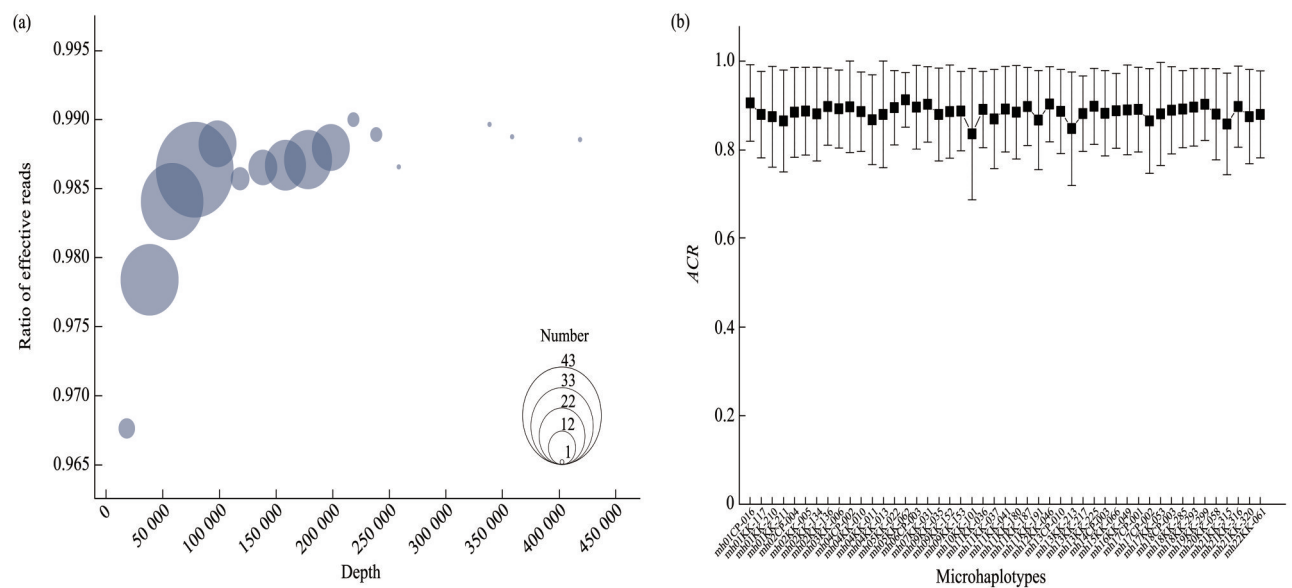


Fig. 1 Evaluation of sequencing data
(a) Bubble chart of the sequencing depth and effective reads ratio of 216 samples. Each bubble represents the number of samples within the sequencing depth range with an increasing series of 20 000; (b) average allele coverage ratio (ACR) distribution of heterozygous loci in 216 samples.

2.2 等位基因频率

在 216 份江苏常熟汉族无关个体的数据中, 共观察到 47 个微单倍型的 239 个不同的等位基因, 不同微单倍型的等位基因数量从 3 个 (mh03KK-006 等 3 个微单倍型, 表 2) 到 13 个 (mh01KK-117) 不等, 各等位基因的频率分布在 0.002 3~0.664 4 之间。在对 1KG 中的 2 504 份无关个体的分型统计中, 47

个微单倍型共观察到 326 个等位基因。通过对比, 47 个微单倍型在本研究数据中有 4 个等位基因在 1KG 的 2 504 份个体数据中未观察到, 分别是: mh04KK-013 的等位基因 24、mh06CP-003 的等位基因 3、mh13KK-213 的等位基因 7, 以及 mh22KK-061 的等位基因 7, 上述等位基因均在杂合子中观察到。

Table 2 Allele frequency distribution of 47 microhaplotypes in 216 unrelated individuals from Changshu, Jiangsu

Microhaplotype	Haplotype	Numeral allele	Frequency	Microhaplotype	Haplotype	Numeral allele	Frequency
mh01CP-016	AGA	3	0.094 9	mh03KK-006	TTC	8	0.305 6
	TAA	5	0.402 8		AA	1	0.594 9
	TGA	7	0.206 0		AG	2	0.307 9
	TGG	8	0.296 3		TA	3	0.097 2
mh01KK-117	AACC	1	0.432 9	mh04CP-002	CAA	5	0.226 9
	AACT	2	0.122 7		CGA	7	0.240 7
	AAGC	3	0.030 1		CGT	8	0.335 7
	AAGT	4	0.011 6		GAA	9	0.196 8
	AGCC	7	0.094 9	mh04KK-010	AA	1	0.664 4
	AGCT	8	0.025 5		AG	2	0.282 4
	AGGC	9	0.002 3		GA	4	0.053 2
	CACC	13	0.185 2	mh04KK-011	AC	2	0.442 1
	CACT	14	0.013 9		AT	3	0.185 2
	CAGC	15	0.030 1		GC	5	0.002 3
	CAGT	16	0.002 3		GT	6	0.370 4
	CGCC	19	0.041 7	mh04KK-013	AAGAT	6	0.104 2
	CGCT	20	0.006 9		CAGAT	22	0.125 0
mh01KK-210	CC	1	0.155 1		CAGGT	24	0.002 3
	TC	3	0.268 5		CGAAT	26	0.002 3
	TT	4	0.576 4		CGGAC	29	0.078 7
mh01KK-211	ACT	2	0.064 8		CGGAT	30	0.113 4
	ATC	3	0.486 1		CGGGT	32	0.574 1
	ATT	4	0.273 2	mh05KK-022	CA	1	0.506 9
	GTC	7	0.175 9		CC	2	0.280 1
mh02CP-004	CGA	5	0.009 3		TC	5	0.213 0
	CGG	6	0.342 6	mh05KK-062	AA	1	0.298 6
	CTA	7	0.453 7		AC	2	0.502 3
	CTG	8	0.094 9		TA	3	0.199 1
	TTG	12	0.099 5	mh06CP-003	AAA	1	0.307 9
mh02KK-005	AG	3	0.381 9		AAC	2	0.215 3
	GA	4	0.365 7		AAT	3	0.004 6
	GG	6	0.252 3		AGC	5	0.314 8
mh02KK-134	ACCG	2	0.037 0		GGC	11	0.157 4
	ACTG	4	0.016 2	mh07KK-031	CA	1	0.500 0
	ATCA	5	0.046 3		CG	2	0.208 3
	ATCG	6	0.495 4		TG	4	0.291 7
	ATTA	7	0.134 3	mh09KK-035	CG	1	0.342 6
	ATTG	8	0.092 6		CT	2	0.435 2
	TCCG	18	0.006 9		TG	3	0.222 2
	TCTG	20	0.013 9	mh09KK-152	AGCA	5	0.226 9
	TTCG	22	0.025 5		ATCG	10	0.002 3
	TTTG	24	0.131 9		ATTA	11	0.074 1
mh02KK-136	GTA	3	0.088 0	mh09KK-153	ATTG	12	0.516 2
	GTC	4	0.083 3		GTCG	22	0.180 6
	TCA	5	0.307 9		CAA	5	0.041 7
	TCC	6	0.215 3		CAC	6	0.074 1

Continued to Table 2							
Microhaplotype	Haplotype	Numeral allele	Frequency	Microhaplotype	Haplotype	Numeral allele	Frequency
mh10KK-101	CGA	7	0.004 6	mh13KK-217	CCG	4	0.148 2
	TAA	9	0.328 7		TAA	7	0.002 3
	TAC	10	0.250 0		TAG	8	0.238 4
	TGA	11	0.224 5		TCA	9	0.273 2
	TGC	12	0.076 4		TCG	10	0.041 7
mh11KK-036	AG	2	0.425 9	mh13KK-225	AACA	1	0.013 9
	CA	3	0.159 7		AACG	2	0.062 5
	CG	4	0.414 4		AATG	8	0.437 5
mh11KK-037	AA	1	0.395 8		AGCA	10	0.134 3
	AG	2	0.245 4		AGCG	11	0.092 6
mh11KK-041	CG	4	0.358 8		AGTA	16	0.004 6
	ACG	1	0.474 5	mh14CP-003	AGTG	17	0.141 2
mh11KK-180	GCG	7	0.349 5		GGCA	28	0.002 3
	GTG	11	0.175 9		GGCG	29	0.094 9
	AG	2	0.053 2		GGTG	35	0.016 2
mh11KK-187	GA	3	0.546 3	mh15KK-066	AAG	2	0.125 0
	GG	4	0.400 5		ACG	4	0.365 7
	AACC	1	0.018 5		GAA	5	0.201 4
	AATC	7	0.060 2		GAG	6	0.294 0
	AATG	8	0.004 6	mh16KK-049	GCG	8	0.013 9
mh11KK-191	ACCC	10	0.497 7		AGG	4	0.101 9
	ACCG	11	0.013 9		GAA	5	0.064 8
	ACTC	16	0.150 5		GAG	6	0.266 2
	ACTG	17	0.023 2		GGA	7	0.430 6
mh12KK-046	GCCC	28	0.009 3		GGG	8	0.136 6
	GCCG	29	0.213 0	mh17CP-001	AG	1	0.500 0
	GCTC	34	0.009 3		AT	2	0.118 1
	CCCA	1	0.525 5		CG	3	0.201 4
	CCCG	2	0.004 6		CT	4	0.180 6
mh13CP-010	GCCG	10	0.011 6	mh17CP-002	AAAAG	2	0.319 4
	GCGA	11	0.006 9		ACAAA	9	0.099 5
	GCGG	12	0.266 2		ACAAG	10	0.006 9
	GTGG	16	0.185 2		ACAGA	11	0.013 9
	CAGT	4	0.243 1	mh17KK-053	ACGGA	15	0.423 6
mh13KK-213	CGAT	6	0.083 3		CCAAA	25	0.136 6
	TAAC	9	0.185 2		CAG	2	0.363 4
	TAAT	10	0.488 4		GAG	6	0.192 1
	GA	3	0.226 9		GGA	7	0.145 8
mh13KK-217	GG	4	0.291 7		GGG	8	0.298 6
	TA	5	0.321 8	mh17KK-053	AGC	1	0.164 4
	TG	6	0.159 7		AGT	2	0.217 6
	AAA	1	0.300 9		GGT	10	0.187 5
	AGA	4	0.083 3		GTT	12	0.430 6
mh13KK-219	AGG	6	0.180 6		CT	2	0.442 1
	GGA	10	0.435 2		TC	3	0.405 1
	CCA	3	0.296 3		TT	4	0.152 8

Continued to Table 2							
Microhaplotype	Haplotype	Numeral allele	Frequency	Microhaplotype	Haplotype	Numeral allele	Frequency
mh18CP-003	ACT	2	0.314 8	mh21KK-315	TGC	7	0.120 4
	ATC	3	0.351 9		ACC	1	0.067 1
	ATT	4	0.238 4		ACT	2	0.004 6
	GTC	7	0.092 6		ATC	3	0.173 6
	GTT	8	0.002 3		ATT	4	0.018 5
mh18KK-285	AGCG	7	0.094 9	mh21KK-316	GCC	5	0.127 3
	CACG	13	0.546 3		GCT	6	0.060 2
	CGCG	19	0.011 6		GTC	7	0.125 0
	CGCT	20	0.134 3		GTT	8	0.423 6
	CGTG	23	0.213 0		ACAC	1	0.416 7
mh18KK-293	AGAA	7	0.261 6	mh21KK-320	ACGC	3	0.004 6
	ATAA	13	0.044 0		ACGT	4	0.324 1
	ATGA	15	0.150 5		ATGC	7	0.067 1
	GGAA	25	0.363 4		GCGC	19	0.187 5
	GGAG	26	0.150 5		AACA	1	0.122 7
	GGGA	27	0.018 5	mh22KK-061	AACG	2	0.273 2
	GTAA	31	0.004 6		GACA	13	0.273 2
	GTAG	32	0.002 3		GACG	14	0.025 5
	GTGA	33	0.004 6		GATA	16	0.185 2
	ATGAA	19	0.120 4		GGCA	19	0.055 6
mh19KK-299	GCAAA	25	0.115 7		GGCG	20	0.064 8
	GCAAG	26	0.423 6		AAA	1	0.192 1
	GCATG	30	0.226 9		AAG	2	0.002 3
	GCGTA	35	0.108 8		GAA	5	0.465 3
	GCGTG	36	0.004 6		GAG	6	0.090 3
mh20KK-058	CAC	1	0.326 4		GGA	7	0.002 3
	TAC	5	0.294 0		GGG	8	0.247 7
	TAT	6	0.259 3				

Compared with the data for 2 504 individuals in 1KG, 4 novel alleles in the Changshu population were shown in bold.

2.3 对比1KG数据的微单倍型多态性分析

通过计算1KG中东亚(EAS)参考人群中5个群体,即CHS群体、CDX群体、越南胡志明市群体(KHV, $n=99$)、CHB群体以及日本东京市群体(JPT)的 A_e 值,216名江苏汉族个体计算的 A_e 与5个群体数据计算的微单倍型的 A_e 值分布均表现出较小的差异(F 检验结果水平显著, P 值均小于0.05,图2a及图S1)。常熟汉族人群的 A_e 作为自变量,CHB群体的 A_e 作为因变量的线性回归分析显示,CHB与常熟汉族人群的预测模型拟合程度最高($R^2=0.903$,图2a及图S1),与CDX群体相关性程度最低($R^2=0.614$,图S1)。另一方面,共18个微单倍型的 A_e 值在江苏汉族人群中高于CHB群体。

基于216名江苏汉族无关个体数据与EAS参考人群中的5个群体数据计算的CMP存在较小差异

(图2b)。得到47个微单倍型在常熟汉族人群中的CMP达到 1.25×10^{-36} 。CHS、CDX、KHV、CHB和JPT群体计算的CMP分别为: 4.51×10^{-38} 、 4.10×10^{-36} 、 6.64×10^{-37} 、 1.41×10^{-37} 和 1.48×10^{-37} 。

2.4 基于1KG的群体遗传分化分析

进一步计算了江苏常熟汉族人群和1KG中5个参考人群,即欧洲人群(EUR)、EAS、美洲人群(AMR)、南亚人群(SAS)和非洲人群(AFR)之间的遗传分化指数,江苏汉族人群与EAS参考人群之间的 F_{st} 值为0.001 6,群体之间存在最小遗传分化($F_{st}<0.05$)。与AFR参考人群之间观察到的分化程度最大 $F_{st}=0.065 7$,两个人群间存在中等遗传分化。这与EAS参考人群和AFR参考人群的结果相似($F_{st}=0.063 7$)。同样地,与1KG中26个群体之间的遗传分化距离显示(图2c),江苏汉族

人群与CHB群体之间的遗传分化指数最低 ($F_{st}=0.006\ 8$)。与AFR参考人群中的Mende in Sierra Leone (MSL) 群体之间达到最大值 ($F_{st}=0.337\ 2$)。

成对Nei氏遗传距离显示(图2d), 26个参考人群及江苏汉族216个个体的数据共形成两个主要分支: AFR参考人群的7个群体形成第一个分支;

AMR参考人群的4个群体、欧洲EUR参考人群的5个群体以及EAS参考人群和SAS参考人群与江苏汉族人群形成第二个分支。在第二个分支中, 江苏汉族人群与CHB群体, CDX群体以及越族EAS参考人群中的KHV群体共同形成一个分支。江苏汉族人群与CHB群体之间的Nei氏遗传距离最小, 为0.006 8。

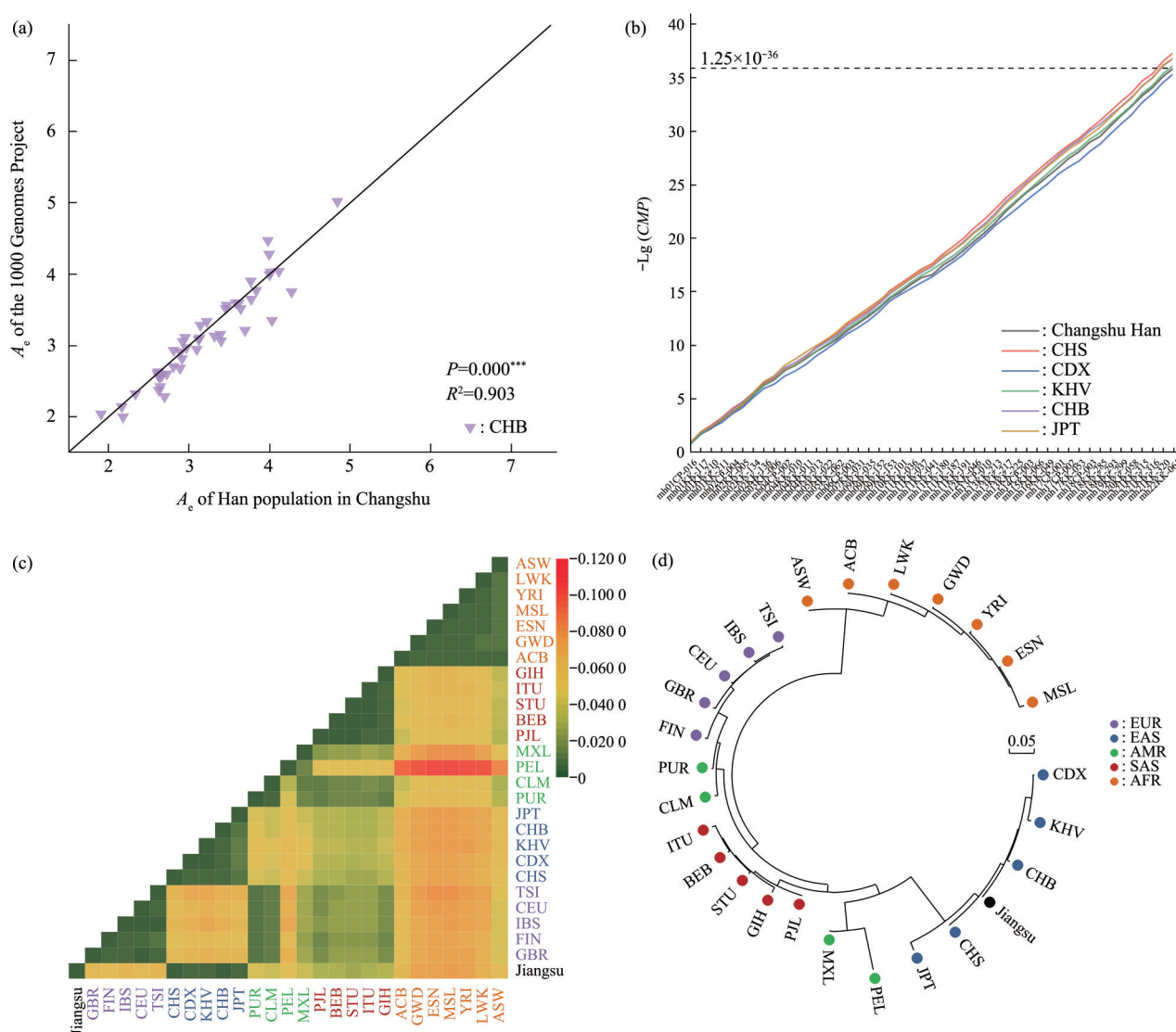


Fig. 2 Comparison of 1000 Genomes Project data and 216 unrelated individuals data in Changshu, Jiangsu

(a) Difference of A_e value. The significant correlation is expressed by the P -value of the F -test is less than 0.05; suffix *** of P -value represents a significant level of 1%. The R^2 value indicates the degree of fit of the regression model, and $R^2=1$ indicates the absolute linear correlation. (b) Difference of cumulative random matching probability (CMP). (c) Fixation index of Changshu population and 26 reference populations of 1000 Genomes Project. (d) The Nei's genetic distance between the Changshu population and 26 reference populations of 1000 Genomes Project. The reference population (or super-population) code in (c) and (d) refers to 1000 Genomes Project^[15].

2.5 法庭科学参数

基于216名常熟汉族无关个体数据计算的47个

微单倍型的基因型分布均符合Hardy-Weinberg平衡 ($P > 0.05$), 所有配对基因座LD检验 P 值范围为

0.004 6~0.933 9 ($P > 0.05$), 矫正值为0.000 2(公式为 $0.5/(n \times (n-1))$), 即微单倍型个数乘以微单倍型个数减去1, 用0.5除以得到的数值, n 为微单倍型个数), 均处于连锁平衡中。表3列出了各个微单倍型的法庭科学参数。47个常染色体微单倍型的 A_c 均处于 1.908 6 (mh04KK-010) 到 4.842 6 (mh21KK-320) 之间, 有6个微单倍型的 A_c 值大于4, 平均 A_c 值为 3.194 4。期望杂合度 (expected

heterozygosity, H_e) 范围为 0.476 0 (mh04KK-010) 到0.793 5 (mh21KK-320), 平均 H_e 为 0.674 5; 个体识别概率 (discrimination power, DP) 在 0.453 6~0.917 0 之间, 平均值为 0.813 9, 非父排除概率 (probability of exclusion, PE) 在 0.160 7~0.609 4 之间, 平均值为 0.390 9, 累积非父排除概率 (combine probability of exclusion, CPE) 为 $1-3.59 \times 10^{-11}$ 。

Table 3 Forensic parameters of 216 unrelated individuals in Changshu, Jiangsu

MH	A_c	DP	PE	TPI	H_o	H_e	$p\text{-HWE}$
mh01CP-016	3.317 1	0.818 0	0.486 3	1.894 7	0.736 1	0.698 5	0.708 0
mh01KK-117	3.995 0	0.895 8	0.558 5	2.250 0	0.777 8	0.749 7	0.159 5
mh01KK-210	2.334 4	0.686 6	0.282 1	1.227 3	0.592 6	0.571 6	0.327 1
mh01KK-211	2.889 6	0.827 0	0.316 0	1.317 1	0.620 4	0.653 9	0.357 1
mh02CP-004	2.922 1	0.825 1	0.287 6	1.241 4	0.597 2	0.657 8	0.018 4
mh02KK-005	2.912 8	0.770 9	0.372 0	1.479 5	0.662 0	0.656 7	0.970 8
mh02KK-134	3.400 6	0.888 5	0.448 6	1.741 9	0.713 0	0.705 9	0.888 6
mh02KK-136	4.013 2	0.888 5	0.558 5	2.250 0	0.777 8	0.750 8	0.533 7
mh03KK-006	2.182 7	0.696 0	0.179 5	0.981 8	0.490 7	0.541 8	0.031 3
mh04CP-002	3.834 5	0.857 8	0.575 2	2.347 8	0.787 0	0.739 2	0.687 0
mh04KK-010	1.908 6	0.650 2	0.160 7	0.939 1	0.467 6	0.476 0	0.227 6
mh04KK-011	2.725 2	0.761 0	0.276 7	1.213 5	0.588 0	0.633 0	0.431 7
mh04KK-013	2.665 9	0.797 0	0.298 7	1.270 6	0.606 5	0.624 9	0.198 5
mh05KK-022	2.626 1	0.733 3	0.346 4	1.402 6	0.643 5	0.619 2	0.887 1
mh05KK-062	2.623 9	0.737 9	0.266 1	1.186 8	0.578 7	0.618 9	0.248 3
mh06CP-003	3.773 1	0.843 2	0.486 3	1.894 7	0.736 1	0.735 0	0.902 2
mh07KK-031	2.642 2	0.786 0	0.316 0	1.317 1	0.620 4	0.621 5	0.409 6
mh09KK-035	2.807 9	0.790 2	0.405 7	1.588 2	0.685 2	0.643 9	0.518 0
mh09KK-152	2.808 8	0.821 3	0.322 0	1.333 3	0.625 0	0.644 0	0.735 3
mh09KK-153	4.272 7	0.879 3	0.566 8	2.297 9	0.782 4	0.766 0	0.710 2
mh10KK-101	2.641 2	0.729 7	0.340 2	1.384 6	0.638 9	0.621 4	0.927 3
mh11KK-036	2.893 3	0.808 0	0.346 4	1.402 6	0.643 5	0.654 4	0.293 9
mh11KK-037	2.643 3	0.775 4	0.304 4	1.285 7	0.611 1	0.621 7	0.185 1
mh11KK-041	2.166 2	0.453 6	0.171 8	0.964 3	0.481 5	0.538 4	0.104 3
mh11KK-180	3.119 4	0.859 8	0.398 8	1.565 2	0.680 6	0.679 4	0.559 8
mh11KK-187	2.621 4	0.766 7	0.310 2	1.301 2	0.615 7	0.618 5	0.246 3
mh11KK-191	2.951 0	0.831 8	0.310 2	1.301 2	0.615 7	0.661 1	0.206 5
mh12KK-046	3.765 5	0.861 8	0.509 8	2.000 0	0.750 0	0.734 4	0.752 8
mh13CP-010	3.130 0	0.839 1	0.340 2	1.384 6	0.638 9	0.680 5	0.362 8
mh13KK-213	4.116 3	0.879 4	0.478 6	1.862 1	0.731 5	0.757 1	0.032 1
mh13KK-217	3.978 7	0.917 0	0.486 3	1.894 7	0.736 1	0.748 7	0.214 2
mh13KK-225	3.615 8	0.867 0	0.525 7	2.076 9	0.759 3	0.723 4	0.503 3
mh14CP-003	3.454 6	0.871 4	0.456 0	1.770 5	0.717 6	0.710 5	0.725 4
mh15KK-066	2.966 5	0.800 9	0.352 7	1.421 1	0.648 1	0.662 9	0.856 9
mh16KK-049	3.222 8	0.839 2	0.456 0	1.770 5	0.717 6	0.689 7	0.535 4
mh17CP-001	3.578 7	0.846 8	0.478 6	1.862 1	0.731 5	0.720 6	0.367 0

Continued to Table 3							
MH	A_e	DP	PE	TPI	H_o	H_e	$p\text{-HWE}$
mh17CP-002	3.391 1	0.865 9	0.412 7	1.611 9	0.689 8	0.705 1	0.986 3
mh17KK-053	2.611 5	0.773 9	0.316 0	1.317 1	0.620 4	0.617 1	0.955 1
mh18CP-003	3.468 2	0.859 2	0.392 0	1.542 9	0.675 9	0.711 7	0.002 0
mh18KK-285	2.695 7	0.729 5	0.328 0	1.350 0	0.629 6	0.629 0	0.222 2
mh18KK-293	4.030 6	0.895 5	0.434 0	1.687 5	0.703 7	0.751 9	0.015 9
mh19KK-299	3.694 8	0.891 6	0.426 8	1.661 5	0.699 1	0.729 3	0.708 0
mh20KK-058	3.640 9	0.862 1	0.398 8	1.565 2	0.680 6	0.725 3	0.087 7
mh21KK-315	4.001 4	0.896 8	0.494 1	1.928 6	0.740 7	0.750 1	0.611 1
mh21KK-316	3.141 5	0.833 4	0.426 8	1.661 5	0.699 1	0.681 7	0.899 2
mh21KK-320	4.842 6	0.905 2	0.609 4	2.571 4	0.805 6	0.793 5	0.491 1
mh22KK-061	3.096 9	0.836 8	0.359 0	1.440 0	0.652 8	0.677 1	0.813 6

MH: microhaplotype; A_e : effective number of alleles; H_e : expected heterozygosity; H_o : observed heterozygosity; DP : discrimination power; PE : probability of exclusion; TPI : typical paternity index; $p\text{-HWE}$: Hardy-Weinberg equilibrium probability.

3 讨 论

目前, MHSeqTyper47 试剂盒已经完成了低质量 DNA 检测中的效能评估^[16, 26]。然而, 微单倍型的筛选和评估更多的是基于人类遗传项目数据^[18, 27-28], 可能会由于缺少个体基因型或测序数据覆盖度低等因素而限制分析。因此, 探索高测序深度下微单倍型在不同人群中的效能将起到重要作用。本研究使用 MHSeqTyper47 试剂盒对江苏常熟地区汉族人群进行微单倍型遗传多态性检测分析, 统计了江苏常熟汉族代表人群频率数据。另外, 对比 1KG 数据, 将为评估实际采集样本与基于数据库的体系理论效能差异提供参考。

早期的研究中统计了低质量样本的测序参数, 在对微量、降解、模拟案件现场样本和抑制剂影响的文库的 3 次测序中, 簇通过率分布在 73.0% 到 90.3% 之间, 3 次测序的碱基质量大于 Q30 的比例分布在 75.0% 到 86.2% 之间^[16]。与本研究检测的人正常唾液卡样对比, 并未观察到测序参数受样本质量的影响。然而, 测序的簇通过率和 Q30 将直接影响测序所得序列数^[24, 29]。本研究中绝大部分样本得到了足够后续分析的序列数(单样本总测序深度在 11 004 以上)。

江苏常熟地区汉族人群 MHSeqTyper47 的 CMP 与 EAS 参考人群中的 5 个群体均结果相近(图 2b), 但明显高于基于 1KG 2 504 名样本代表的全球水平的 8.70×10^{-44} 。这可能是因为受人群之间的特有等位基因影响, 比较 5 个参考人群分别的 CMP 可以证实这一现象, 如基于 5 个参考人群 EUR、

EAS、AMR、SAS 和 AFR 数据计算所得 CMP 分别为 2.04×10^{-37} 、 1.64×10^{-38} 、 1.58×10^{-39} 、 1.97×10^{-40} 和 1.45×10^{-37} 。常熟地区汉族人群 A_e 值与 1KG 中 EAS 参考人群的 5 个群体间并没有显著的差异, 但结果表明, 常熟汉族人群的 A_e 分布与 CHB 群体之间相似性最高, 这将说明, CHB 群体更好的代表常熟汉族人群的多态性。另外, 对比现有商业化 STR 试剂盒, 47 个微单倍型在常熟地区 216 名汉族无关个体的 CMP 低出 Goldeneye™ 20 A 试剂盒 19 个常染色体 STR 在江苏地区 10 000 名个体中的数值(1.16×10^{-23})^[30]。

为了评估 1KG 中 EAS 参考人群与江苏常熟汉族人群的遗传分化程度, 成对的 F_{st} 显示: 常熟汉族人群与 EAS 参考人群较为相似, 而与 AFR 和 EUR 参考人群存在较大的遗传分化。群体间 Nei 氏遗传距离也表现出相似的结果。综上所述, 相比于 1KG 全部群体数据, 基于 1KG 中 EAS 参考人群数据进行微单倍型筛选或评估更符合中国人群遗传特征。

4 结 论

本研究报告了 47 个微单倍型基因座在江苏常熟地区汉族人群中的等位基因频率及遗传多态性信息, 为 47 个微单倍型在法医学应用中提供了数据基础。另外, 比较了 1KG 参考人群与江苏常熟地区汉族人群的多态性差异, 并揭示了 47 个微单倍型在江苏常熟地区汉族人群中的遗传结构。总的来说, 1KG 中的东亚人群参考数据更符合江苏常熟地区汉族人群的遗传特征。

附件 见本文网络版 (<http://www.pibb.ac.cn> 或 <http://www.cnki.net>):

PIBB_20230027_Figure_S1.pdf

参 考 文 献

- [1] Kidd K K, Pakstis A J, Speed W C, *et al.* Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. *Forensic Sci Int Genet*, 2014, **12**: 215-224
- [2] 饶旻, 李彩霞, 赵钊, 等. 微单倍型遗传标记及其法医遗传学应用. *刑事技术*, 2017, **42**(4): 324-328
Rao M, Li C X, Zhao Z, *et al.* *Forensic Sci Technol*, 2017, **42**(4): 324-328
- [3] 于文帅, 康克莱, 张驰, 等. 微单倍型遗传标记研究新进展及其在法医混合 DNA 分析中的应用. *沈阳药科大学学报*, 2023, **40**(9): 1245-1252
Yu W S, Kang K L, Zhang C, *et al.* *Journal of Shenyang Pharmaceutical University*, 2023, **40**(9): 1245-1252
- [4] Yu W S, Feng Y S, Kang K L, *et al.* Screening of highly discriminative microhaplotype markers for individual identification and mixture deconvolution in East Asian populations. *Forensic Sci Int Genet*, 2022, **59**: 102720
- [5] Fan H, Xie Q, Wang L, *et al.* Microhaplotype and Y-SNP/STR (MY): a novel MPS-based system for genotype pattern recognition in two-person DNA mixtures. *Forensic Sci Int Genet*, 2022, **59**: 102705
- [6] Hiroaki N, Koji F, Tetsushi K, *et al.* Approaches for identifying multiple-SNP haplotype blocks for use in human identification. *Legal Med*, 2015, **17**(5): 415-420
- [7] Yang J, Chen J, Ji Q, *et al.* A highly polymorphic panel of 40-plex microhaplotypes for the Chinese Han population and its application in estimating the number of contributors in DNA mixtures. *Forensic Sci Int Genet*, 2022, **56**: 102600
- [8] Zou X, He G, Liu J, *et al.* Screening and selection of 21 novel microhaplotype markers for ancestry inference in ten Chinese subpopulations. *Forensic Sci Int Genet*, 2022, **58**: 102687
- [9] Phillips C, Mcnevin D, Kidd K K, *et al.* MAPlex - a massively parallel sequencing ancestry analysis multiplex for Asia-Pacific populations. *Forensic Sci Int Genet*, 2019, **42**: 213-226
- [10] Bennett L, Oldoni F, Long K, *et al.* Mixture deconvolution by massively parallel sequencing of microhaplotypes. *Int J Legal Med*, 2019, **133**(3): 719-729
- [11] Qu N, Lin S, Gao Y, *et al.* A microhap panel for kinship analysis through massively parallel sequencing technology. *Electrophoresis*, 2020, **41**(3-4): 246-253
- [12] Wu R, Li H, Li R, *et al.* Identification and sequencing of 59 highly polymorphic microhaplotypes for analysis of DNA mixtures. *Int J Legal Med*, 2021, **135**(4): 1137-1149
- [13] Pang J B, Rao M, Chen Q F, *et al.* A 124-plex microhaplotype panel based on next-generation sequencing developed for forensic applications. *Sci Rep*, 2020, **10**(1): 1945
- [14] Sudmant P H, Rausch T, Gardner E J, *et al.* An integrated map of structural variation in 2, 504 human genomes. *Nature*, 2015, **526**(7571): 75-81
- [15] Consortium T G P. A global reference for human genetic variation. *Nature*, 2015, **526**(7571): 68-74
- [16] Feng Y S, Zhang C, Chen Q F, *et al.* Evaluation of the MHSeqTyper47 kit for forensically challenging DNA samples. *Forensic Sci Int Genet*, 2022, **61**: 102763
- [17] Kidd K K, Speed W C, Pakstis A J, *et al.* Evaluating 130 microhaplotypes across a global set of 83 populations. *Forensic Sci Int Genet*, 2017, **29**: 29-37
- [18] Chen P, Yin C, Li Z, *et al.* Evaluation of the microhaplotypes panel for DNA mixture analyses. *Forensic Sci Int Genet*, 2018, **35**: 149-155
- [19] Chen P, Deng C, Li Z, *et al.* A microhaplotypes panel for massively parallel sequencing analysis of DNA mixtures. *Forensic Sci Int Genet*, 2019, **40**: 140-149
- [20] Zhang C, Cao Y D, Song J J, *et al.* MHTyper: a microhaplotype allele-calling pipeline for use with next generation sequencing data. *Aust J Forensic Sci*, 2021, **53**(3): 283-290
- [21] Song J J, Zhang C, Kang K L, *et al.* Numeral nomenclature proposed for microhaplotype alleles to exert efficient forensic applications. *Forensic Sci Technol*, 2022, **47**(6): 647-651
- [22] Excoffier L, Lischer H E L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resources*, 2010, **10**(3): 564-567
- [23] Guo J, Liu Y, Peng Y, *et al.* Genetic polymorphism of 21 non-CODIS STR loci for Han population in Hunan Province, China. *Forensic Sci Int Genet*, 2015, **17**: 81-82
- [24] Churchill J D, Schmedes S E, King J L, *et al.* Evaluation of the Illumina® beta version ForenSeq™ DNA signature prep kit for use in genetic profiling. *Forensic Sci Int Genet*, 2016, **20**: 20-29
- [25] Yang K R, Miao L, Kang K L, *et al.* Sequence polymorphisms of forensic Y-STRs revealed by a 68-plex in-house massively parallel sequencing panel. *Forensic Sci Int Genet*, 2022, **59**: 102727
- [26] 王鑫, 崔扬, 陈维忠, 等. 微单倍型遗传标记在法医物证检验中的应用. *中国法医学杂志*, 2022, **37**(4): 362-364+368
Wang X, Cui Y, Chen W Z, *et al.* *Chinese Journal of Forensic Medicine*, 2022, **37**(4): 362-364+368
- [27] Wu R, Li H, Li R, *et al.* Identification and sequencing of 59 highly polymorphic microhaplotypes for analysis of DNA mixtures. *Int J Legal Med*, 2021, **135**(4): 1137-1149
- [28] Van Der Gaag K J, De Leeuw R H, Laros J F L, *et al.* Short hypervariable microhaplotypes: a novel set of very short high discriminating power loci without stutter artefacts. *Forensic Sci Int Genet*, 2018, **35**: 169-175
- [29] Jager A C, Alvarez M L, Davis C P, *et al.* Developmental validation of the MiSeq FGx forensic genomics system for targeted next generation sequencing in forensic DNA casework and database laboratories. *Forensic Sci Int Genet*, 2017, **28**: 52-70
- [30] Pan M, Ye Q, Ju X B, *et al.* Population genetic polymorphism and mutation analysis of 19 autosomal STR loci in Jiangsu Han individuals in Eastern China. *Ann Human Biol*, 2019, **46**(3): 272-278

The Genetic Polymorphism and Structural Analysis of 47 Microhaplotypes in a Jiangsu Changshu Chinese Han Population*

PAN Kun-Peng^{1,2)**}, FENG Yao-Sen^{2)**}, YU Wen-Shuai^{2,3)}, LIU Zong-Wei⁴⁾, YAO Yi-Ren²⁾, ZHAO Jie²⁾, KANG Ke-Lai²⁾, ZHANG Chi²⁾, WANG Le^{2)***}, WU Jian^{1)***}

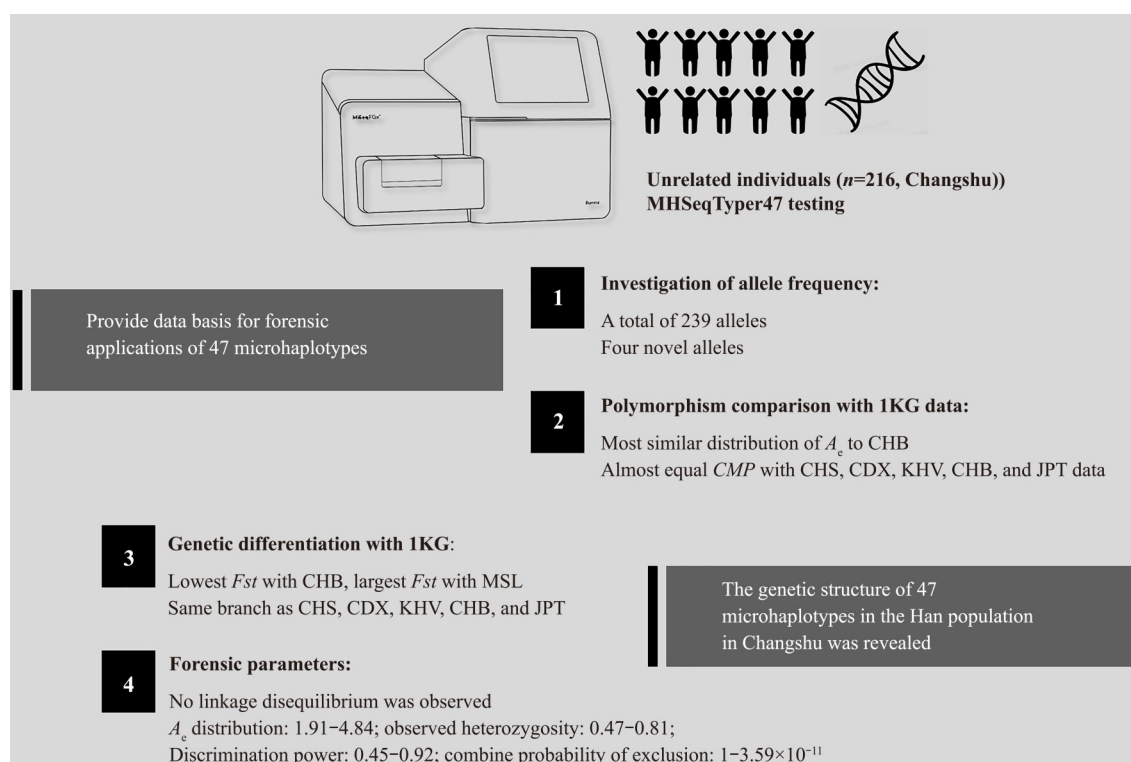
(¹⁾School of Forensic Medicine, Kunming Medical University, Kunming 650500, China;

²⁾National Engineering Laboratory for Forensic Science, Key Laboratory of Forensic Genetics of Ministry of Public Security, Institute of Forensic Science, Ministry of Public Security, Beijing 100038, China;

³⁾College of Investigation, People's Public Security University of China, Beijing 100038, China;

⁴⁾Changshu Public Security Bureau, Changshu 215500, China)

Graphical abstract



* This work was supported by grants from The National Natural Science Foundation of China (81971797), the National Key R&D Program of China (2022YFC3341001), and the Ministry of Public Security of China (2017JSYJA05).

** These authors contributed equally to this work.

*** Corresponding author.

WANG Le. Tel: 86-10-83752707, E-mail: wangle_02@163.com

WU Jian. Tel: 86-871-65922837, E-mail: wujianky@126.com

Received: January 21, 2023 Accepted: April 18, 2023

Abstract Objective To investigate the genetic polymorphism and structure of 47 autosomal microhaplotypes in the Han population in Changshu City, Jiangsu Province, and to evaluate the forensic efficiencies and forensic parameters. **Methods** The DNA library of unrelated individual samples was prepared according to MHSeqTyper47 kit manual and sequenced on the MiSeq FGx platform. Microhaplotype genotyping and sequencing depth statistics were processed using MHTyper. The genetic information of samples was then evaluated. The fixation index and genetic distance between the Jiangsu Changshu population and the reference populations in the 1000 Genomes Project phase 3 (1KG) were calculated, and forensic parameters were evaluated. **Results** The fixation index and genetic distance between the Han population in Changshu, Jiangsu, and the CHB (Han Chinese in Beijing, China) reference population in 1KG were the lowest. The effective allele number (A_e) of each locus is also the closest between the two populations. The combined matching probability (CMP) of the Changshu Han population is close to the 5 populations of the East Asian reference super-population in 1KG, which is 1.25×10^{-36} , and the combined probability of exclusion reached 0.999 999 999 964 1. **Conclusion** This study reported the genetic polymorphism and allele frequency of 47 microhaplotypes in a Han population in Changshu City, Jiangsu Province. This information provides a data basis for 47 microhaplotypes in forensic applications. In addition, the polymorphism differences between the 1KG reference population and the Han population in Changshu, Jiangsu were compared, and the genetic structure of 47 microhaplotypes in the Han population in Changshu, Jiangsu was revealed. In general, the reference data of the East Asian super-population in 1KG is more in line with the genetic characteristics of Han population in Changshu, Jiangsu.

Key words forensic genetics, microhaplotype, next generation sequencing

DOI: 10.16476/j.pibb.2023.0027